INSTITUTE of Museum and Library SERVICES

FINAL PERFORMANCE REPORT

For Projects with Award Dates before September 30, 2015

For instructions, please see Guidance for Preparing and Submitting Your Final Performance Report Package.

1. Federal agency and organization element to which report is submitted: **Institute of Museum and Library Services**	2. Federal award or other identifying number assigned by federal agency: LG-71-15-0174	Page 1	of 11 Pages
		3a. D-U-N-S® number: 059453410	
		3b. EIN/TIN: 94-3242767	
4. Recipient organization (name and complete address, including ZIP+4/postal code): Internet Archive 300 Funston Avenue San Francisco, CA 94118-2116		5. Recipient identifying or account number: LG-71-15-0174	
6a. Award period of performance start date (MM/DD/YYYY): 01-01-2016	6b. Award period of performance end date (MM/DD/YYYY): 12-31-2018	7. Reporting period end date (MM/DD/YYYY): 12-31-2018	
8. Project URLs, if any: https://github.com/WASAPI-Community/data-transfer-apis https://archive.org/details/wasapi		9. Report frequency: ☐ annual ☐ semi-annual ☐ quarterly ☐ other If other, describe:	
10. Other attachments? ☐ Yes ☐ No Contact the appropriate IMLS program office to receive instructions for transmitting additional attachments.			
11a. Name and title of Project Director: Jefferson Bailey Director, Web Archiving & Data Services		11b. Telephone (area code, number, extension): 1-415-561-6767	
		11c. Email address: jefferson@archive.org	
12. Certification: By submitting this report I certify to the best of my knowledge and belief that this information is correct and complete for performance of activities for the purposes set forth in the award documents.			
13a. Signature of Authorized Certifying Official:		13b. Date report submitted (MM/DD/YYYY): 04/16/2019	
13c. Name and title of Authorized Certifying Official: Jefferson Bailey Director, Web Archiving & Data Services		13d. Telephone (area code, number, extension): 1-415-561-6767	
		13e. Email address: jefferson@archive.org	

Final Performance Report Narrative, Year Three
Identifying Number: LG-71-15-0174; Recipient: Internet Archive
Grant Period: 01/01/2016 to 12/31/2018; Reporting Period: 01/01/2018 to 12/31/2018

Project Title: Systems Interoperability and Collaborative Development for Web Archiving
Submitted By: Jefferson Bailey, Director, Web Archiving & Data Services, Internet Archive (PI)

Description of Project Partners

- Internet Archive: lead institution, developed API specification, built and implemented WASAPI API for use by institutions in the Archive-It service and conducted user testing, training, documentation, outreach, etc. IA convened the project's National Symposium, lead all R&D, and was author/co-author and publisher of all project research reports.
- Stanford Libraries / LOCKSS Program: SUL/LOCKSS implemented use of the WASAPI API for data transfer into their local preservation repository, co-developed the data model, conducted project-related training and outreach, and contributed technical development and expertise and authored/co-authored the project's research reports.
- University of North Texas: UNT contributed to the design specifications for the transfer API, developed a custom, Python-based client for working with the API, and contributed to project outreach and communications.
- Rutgers University: The team at Rutgers used the WASAPI API for researcher dataset request, creation, and transfer in the course of doing computational analysis of their web archive collections. They also contributed to data modeling and API design and project outreach and communication, including notable publications on their work in professional journals and conferences for data-driven social science research.

Overview

The Systems Interoperability and Collaborative Development for Web Archiving grant (codenamed WASAPI: Web Archiving System APIs) was a research and development project to address the need for social and technical systems to expand the community of technical contributors to web archiving software and to facilitate the interoperability of web archiving systems for the sake of enhanced preservation and access for archived web data. The audience was digital librarians, archivists, technologists, and computational researchers. As an R&D project featuring extensive in-production software releases and API development, as well as community building, the project was successful. The WASAPI project advanced interoperability between critical pieces of archiving technical infrastructure; outlined the needs for a more inclusive, empowered, contributory community of library technologists; and built out social and technical architecture for further work to integrate currently siloed web archiving and digital library technology systems.

Changes

The key change was that Year Three was a one-year, no-cost extension of the grant to allow for a Stakeholders Meeting and to finalize the project's reports for publication in Q2 2019 (as well as spend the small portion of remaining funds). The only notable events in this regard were the expansion of scope to include the work of overlapping projects within the grant partners' own institutions (no funding was requested for this addition). This includes documenting affiliate APIs by Internet Archive that complement the WASAPI work, additional collaboration between project partners and downstream services and users of the project's APIs, and further outreach and community to disseminate the project's outcomes and establish WASAPI as the hub of API and web archiving systems interoperability collaboration going forward.

Activities Completed During This Reporting Period

During Year Three of the project *Systems Interoperability and Collaborative Development for Web Archiving*, all main activities to be completed were successfully accomplished. As this project was originally a two-year grant, Year Three consisted of a one-year, no-cost extension to enhance project deliverables through additional community building and knowledge sharing activities supported by leftover grant funds originally earmarked for attendee travel to the project's National Symposium. Thus, the scope of activities completed in Year Three were minor in comparison to prior years and the annual reports for Year One and Year Two are more illustrative of overall project work and provide more extensive discussion of outcomes accomplished in those stages. Overall themes from those reports are reiterated, in brief, in this report for summation and to link them to the activities of Year Three. The project continued operating under the working name of the "WASAPI" project (Web Archiving Systems APIs), for clear identification, community building, and to establish and proselytize a foundational network of contributing and participating institutions that can continue to work together beyond the grant period. At a high level, Year Three work had the following outcomes:

- A WASAPI Stakeholders Meeting, held at the offices of CLIR in Washington D.C. and featuring project staff and 10 additional collaborators and experts. Outcomes from this one-day meeting were both better coordination amongst different projects utilizing WASAPI-based APIs for data transfer and a roadmap of future directions for continuing the project's work.
- Final composition and editing of the project's formal public publications. These documents have been drafted but are undergoing final author edits and are scheduled for publication in May of 2019, prior to the annual convening of the International Internet Preservation Consortium meeting in June 2019. The research papers include:
    a. White paper on API-based interoperability for web archives,
    b. Summary report from the project's Year Two National Symposium, and
    c. Summary report from the project's Year Three Stakeholders Meeting.

- Since Year Three was a no-cost extension year to spend the leftover grant funds (~$10,000), there were no additional deliverables specific to the original grant, other than the Stakeholders Meeting and the publication of the project's research papers. However, as WASAPI-related work was intended to extend well beyond the project's original two-year grant timeline, continued activities, advancements, and deliverables we achieved. These include:
    a. 4 additional conference/community presentations, forums, or discussion groups delivered including at IIPC, CNI, SAA, and others, as well as webinars and video presentations for multiple professional groups.
    b. Migration of a majority of Archive-It institutional users to using the WASAPI API for data transfer of their web archives to their local repositories.
    c. Multiple research projects were initiated between Archive-It and computational researchers that leveraged the WASAPI API for requesting and transfer for research datasets.[1]
    d. The collaboration between the Archives Unleashed project and Archive-It was extended and formalized, with nearly 200TB of data transferred via WASAPI for data mining of Archive-It collection as part of the Archives Unleashed Cloud service and as part of their series of "datathon" events training librarians and researches in the tools and methodologies of working with web archive data as scale for scholarly analysis.

Project Results

Since the main activity of the no-cost Year Three work was the convening of a Stakeholders Meeting, it merits summarizing the takeaways explicated in the forthcoming report of that meeting that will be published as the [working title] document, "Insights on Fostering a Collaborative Technical Community for Web Archiving." The summary falls under the three main research areas of the overall WASAPI project.

1. *What are the attributes of a community model that can support sustainable and broad-based collaborative web archiving technology development?*
    a. Project Successes: The project's National Symposium demonstrated broad interest in a U.S.-based convening focusing on preserving the web and fostering technical collaboration. Similarly, adoption of the project's technical work successfully

[1] Discussion of some of these project related to analyzing local news can be seen at, https://www.cjr.org/tow_center_reports/the-dire-state-of-news-archiving-in-the-digital-age.php and, most notably, the work of the News Measures Research Project, https://dewitt.sanford.duke.edu/wp-content/uploads/2018/08/Assessing-Local-Journalism_100-Communities.pdf.

demonstrated the potential and utility of API-based interoperability between services across the web archiving landscape.

b. Project Challenges: While there is broad interest in an annual meeting per the above, few institutions have the staff, time, or money to provide an institutional/administrative home for such a convening. Membership fatigue will likely make a consortial, independent host unlikely. Affiliated conferences or meetings are themselves too money or capacity strapped to add web archiving to already crowded agendas. Similarly, technical capacity at most libraries doing web archiving remain small or nonexistent. A successful model is more likely build around extending, scaling the few instead of bootstrapping the many.
c. Potential for Future Work: The WASAPI project team continues to explore partnerships for an North American annual event on web archiving technologies.

2. *What are the community needs and possibilities for the planned open API to facilitate transfer of web archive data between distributed systems and what other prospective APIs does it point to?*
    a. Project Successes: The project designed, built, and implemented multiple production APIs for data transfer now being used by dozens of institutions for hundreds-of-terabytes scale data distribution and preservation. Multiple utilities were also developed for local use of the WASAPI spec/API for ingest of archived web data into local repositories. Follow-on APIs were modeled, and, in some cases such as Archive-It, implemented because of the project's work.
    b. Project Challenges: While the data distribution between interoperable systems was a success, the overall extent of archived web data that was replicated to other systems only fractionally increased. Instead, a large portion of data distribution was for computational analysis and research services. An excellent outcome, indeed, and in scope for the grant's objectives, but this was seen as a stretch goal. That data reuse and research was the area primarily benefiting from WASAPI APIs was a surprising, if pleasant, outcome of the project's research.
    c. Potential for Future Work: The WASAPI project has laid the groundwork for additional investment in connecting web archiving systems to support data-driven research. Internet Archive is in active talks with data mining projects to build on WASAPIs success in supporting researchers in analyzing web archives and expects new partnerships and services to emerge due to WASAPI's success.

3. *How can better interoperability of web archiving systems support new forms of access and research use?*
    a. Project Successes: Facilitating research use of archived web data was perhaps one of the greatest unplanned successes of the WASAPI project. The project partners

at Rutgers University conducted a scholarly research project data mining archived community newspapers that received national coverage. As well, the Archives Unleashed project received significant funding to build technologies, host events, and build community for computational research uses of web archives with its core systems designed from scratch to interoperate with WASAPI APIs. Having included research dataset support services into the API implementation at Archive-It (a feature beyond the original grant deliverables) allowed research partnerships to extend beyond the expectations of the project team.

b. Project Challenges: There were few challenges to this aspect of the project's work, since it succeeded beyond expectations; however, business/sustainability models for computational research services need further development and testing.
c. Potential for Future Work: Other services beyond the project institutions are in the process of implementing the WASAPI API specification to enable archived web data transfer for research analysis. The Archive-It team plans to extend the research datasets enabled by the WASAPI to build upon the grant's success.

Beyond the Year Three work, the project results were report in the Year Two Interim report and are reiterated here in summary for the sake of final reporting. Readers are encouraged to review the Year Two Interim Report for more details discussion of these results.[2] There were two primary areas of work across the grant partners in Year Two that advanced this strategic project goal. The first was **social**, primarily consisting of the convening of the National Symposium, which both provided a forum for community members to detail their local uses of existing web archiving APIs and a chance to articulate and document the many use cases for which the WASAPI project and future API work could improve existing workflows and new processes and functionalities. In addition, focused conversations at many of the outreach and promotion events, and inclusion of continued information-seeking questions in annual survey mechanisms by NDSA and Archive-It, helped advance this work. The second was **technical**, consisting primarily of engineering, testing, and iterative improvement of the project's suite of APIs and utilities for web archive data transfer to local repositories and for downstream use.

The **social** work elucidated a number of community needs, including:

- Need for more robust tools with better metadata for downloading web archive data from service providers to local preservation repositories.
- Need for better access to externally-stored WARC files for use in creating custom local discovery systems via local indexing.

[2] https://archive.org/details/wasapi

- Need for APIs to facilitate multi-institutional aggregated collections with federated search and index building and the need to enable easier inter-institutional transfer of web archive data to avoid duplicative activities but enable distributive custodialism.
- Need for APIs upon which to build second-level services, such as quality assurance, capture improvement, format migrations (or preservation actions in general), ingest management, data analytics, indexing, et cetera.
- Need for transfer and local processing of WARC files for enabling research and data mining use cases, such as powering Jupyter Notebooks or data extraction.

Discussion on prospective future WASAPI APIs beyond this grant fell into two categories:

- **Metadata APIs**: Though "metadata" is a vague term in web archiving (where technical and administrative metadata tend to endlessly proliferate), the most common inference of the term in the context of these conversations was for "descriptive" and "seed" metadata. This meant, broadly, user-applied descriptive metadata such as subject headings, collection descriptions, or other human classification metadata and seed metadata meaning, mostly, settings applied to specific seeds or URLs being archived, such as frequency of capture, scoping rules, and other acquisition management parameters.
- **Crawl Configuration APIs**: These prospective APIs included information on more granular aspects of crawling configuration, such as robots.txt handling, seed type information (e.g., archiving one URL, one URL plus embedded content, etc.), and capture limits or expansions, such as exclusions of files over a specific size, host-level rules, automated behavior or other utilities used by the crawler, et cetera.

The **technical** work involved the building, testing, and iterative improvement of the APIs developed as part of this project. This elucidated a number of unforeseen issues and possibilities:

- Temporal complexity: Any web archiving program is running numerous crawls at various frequencies across many collections. While a start-date and end-date can be associated with a specific crawl job, and all WARCs have an associated timestamp at which the file was created, exploration of use cases quickly exposes an extensive set of complexities associated with web archiving that can prove challenging to building data models supporting APIs. A crawl job may not include all the seeds in a collection, so crawl job and collection are not 1-to-1 metadata points. To address this, the Archive-It API included the ability to query WARCs by separate data ranges associated with the crawl job and with the WARC creation date itself, i.e. "give me all the WARCs associated with all the crawls within this time range" or "give me all the WARCs from this time range even if part of a crawl job time range falls outside of it."
- Packaging and formats: WARC files (the ISO standard for web archives) are rather unique in that they conventionally represent the crawling process, not a specific information package like a document or an image. A WARC can contain millions of tiny

files or just one big file. Supporting transfer APIs based specifically on data potentially consisting of only parts of WARC files requires extracting that data from existing WARCs and writing that data to new WARC files. This is too complex and computationally expensive for most use cases. The Archive-It API addressed this issue by introducing a data point in the APIs response that indicates if the set of results from a query "includes-extra" – meaning, this list of WARCs includes the data requested by an API request, but also includes other data within these WARCs that does not conform to a given query, such as a seed (whose data may only fill a portion of a WARC file).

- Crawl granularity: As suggested above, associating a specific "crawl job" with a set of seed URLs, configurations, timestamps, and other technical metadata points can be complex. As a consequence, API query parameters based on crawl/job IDs may potentially only be applicable to some portions of a total collection. Therefore, web archive data transfer APIs need to enable query parameters that account for the incompleteness of other query parameters.
- Results delivery: Collections may contain tens or hundreds of thousands of WARC files. Delivering extensive metadata about each individual file in such a collection introduces performance challenges, given the volume of data returned for a query. Thus an API needs to support unexpected functionality such as pagination. In addition, to support the research use cases outlined above, the API needed to also return results related to derivative files, such as CDX, WAT, or WANE files. But since these files are not preservation formats (since they can always be regenerated from WARCs, they need not be preserved long-term) their existence in results may be ephemeral, since these results contain links to on-disk locations which, for derivative files, may only be temporary.

While some of the points outlined here may delve into technical minutiae, they illustrate the broader challenges of pairing user expectations around features and functionality and the technical contingencies of building APIs based on tools and processes that are often obscure, complex, or unknown to archive managers or downstream users. Illuminating and exploring some of these challenges was a specific mandate of this research and development project, and many of these issues only emerged through technical development, not through requirements gathering or scenario planning, so better understanding of these issues can be considered a successful outcome of the technical work of the project.

Though funded as in the "Research" project category, the grant proposal included significant engineering and development work intended both for R&D purposes, but also to build sustainable, at-scale, in-production systems. Thus, the project's scope of work features the requirements development, planning, build out, testing, iterative improvement, and production release of a number of data transfer APIs. All these APIs are open-source and are documented and can be used, forked, and improved by the broader community.

The Archive-It data transfer API is fully documented in the WASAPI project's Github repository (https://github.com/WASAPI-Community/data-transfer-apis). This API is already in use by dozens of Archive-It partners who actively use it to transfer their data collected using Archive-It into their own local preservation repositories. The API is fully documented and has already led many partners to build local tools to interact with the API. The build-out of this API included features beyond the scope of the grant work, including the ability to submit a job to request the creation of research datasets and the ability for the API to notify a requestor when a job is completed and then provide results via the API to the information related to those research datasets (whose data model generally conforms to those of the corresponding preservation file) -- essentially enabling a research-datasets-as-a-service feature as part of the transfer API.

The Mellon-funded LOCKSS software re-architecture will result in a LOCKSS system re-built around API-based web service components and an underlying WARC data store (regardless of the original format of the source content). The WASAPI data transfer APIs provide a common and predictable interface for the export of content out of a LOCKSS system. With the foundation of a common format and APIs, consumers of data from LOCKSS systems can capitalize on the growing array of tools, workflows, and expertise around WARC-stored content and WARC derivative formats. The University of North Texas has also built an open-source local utility and tested it with production transfers of web data between institutions. Similarly, Stanford Libraries continues to invest in automation to streamline the flow of web archives created using Archive-It into local preservation, access, and discovery systems. The WASAPI data transfer APIs allow for easier and more robust integration of this hybrid and distributed infrastructure, freeing up developer time to focus on where they can add particular value to their web archive collections, through enhanced and integrated local access and discovery services. Rutgers University has built a tool for researchers that uses the WASAPI API to find files, request datasets, download datasets, run local data mining and data analysis scripts, and generate a data visualization output.

Lastly, an unexpected, but exciting, occurrence over the course of the project's Year Two work, was the start of development of second-level services by non-grant-partner institutions using the project's production APIs. The Archives Unleashed project (http://archivesunleashed.org/) has built a researcher workbench platform that utilizes the WASAPI API for transferring data from Archive-It collections to their infrastructure for local processing to support research use of collections. Other projects, such as Webrecorder and Islandora have tested the WASAPI APIs for exploring system and service integrations. Supporting such community-wide efforts towards interoperability across tools and services was, of course, one of the main goals of this grant; however that such work began even during the grant period itself, instead of afterwards, was both a pleasant surprise but also a direct affirmation of the grant original statement of need and

argumentation. Supporting, extending, and scaling this ecosystem of interoperability across web archiving systems will continue in Year Three and beyond.

What's next?

Community Models for Collaborative Web Archiving Technology Development

The National Symposium demonstrated the need for a dedicated, national-level web archiving annual meeting exists in the community. Current technical work is fractured across institutions and minimally documented or promoted. Local use cases may vary, and the suite of web archiving tools benefiting from collaborative development is large, but there is community willingness to invest in the core technologies and combine them with other digital library tools to achieve workflow and resource efficiencies and advance the quality and scale of our collective efforts to collect and preserve historical records published on the web. Fostering this social architecture in an environment of membership fatigue, conference bloat, and limited local capacity for the administration and event planning burden of running such a recurring meeting remains a challenge. While national-level conferences were held at University of Michigan, Columbia University, and Internet Archive in the past five years, these institutions are not well suited to take a leadership role in running a recurring yearly meeting. At the same time, IIPC remains a consortial entity with primarily European members and primarily National Libraries represented. Given the differences in resources and mandates between National Libraries and all but the largest research universities, IIPC remains an insufficient venue for many U.S.-based institutions doing web archiving at a smaller scale on fractional staffing.

Lastly, the question remains whether institutions find local value in contributing technical resources to an area of library activity (web archiving) that is almost entirely outsourced to vendors, is a miniscule portion of the library acquisition budget, and that, in most libraries, is, at best, fractionally staffed. Unlike technical contributions to core repository or ILS software projects, the return-on-investment in contributing to web archiving technology development is likely to strike administrators as small. Thus, a community model will need to account for participation of a small number of highly-skilled, highly-knowledgeable technologists, not a large number of minor contributors as originally conceptualized.

Needs and Requirements for Web Archive Data Transfer and Other APIs

Development, iteration, and production deployment of numerous transfer APIs, local utilities, and research data mining tools provided ample insight into the complexities and possibilities for further interoperability between web archiving and digital library systems, as well as how these can inform research use cases. Though available people resources for development on web archiving tools remains limited, benefits of scale are possible when well organized through structured community building, especially with researchers and computational users not coming to archival web data from the library. Modularity in architecture, efficiency in performance,

flexibility in data modeling, and other characteristics will prove essential in planning future engineering efforts support WASAPI's goals, as proven by project work. In some ways, the API-based data transfer, for preservation replication and for research use cases, was the most successful aspect of the project's R&D, with many institutions migrating to using Archive-It's API for data transfer, downstream services like Archives Unleashed using it for research support services, and adoption of the API as standard for interoperability between systems (other web archiving services are working to implement WASAPI-based transfer mechanisms).

Potential for new Web Archiving APIs to Support New Technologies and Uses
As noted above, grant successes include unforseen system integrations already emerging during the grant's research and development phase, including preservation services, data mining applications, harvesting tools, access interfaces, and more. This points to a significant potential for connecting a variety of systems that can support all aspects of the web archiving lifecycle, not just those targeted in this project. This potential "systems interoperability landscape" is better understood, operationally and technically, and is maturing, as a result of the grant's work. Additionally, data mining tools and service are now in full production deployment using the project's technical outputs and are already supporting the work of researchers in their analytical work and appearing in presentations and scholarly publications. As well, adoption of the WASAPI project's work has led project lead institutions to advance work on affiliated APIs for further interoperability of systems around web collection metadata and crawling configurations. The grant's success so far has enabled both these trends and has initiated a community-wide effort to sustain the WASAPI accomplishment in advancing web archiving and libraries' ability to preserve and provide access to our nation's historical record in born-digital form.

Specific ongoing activities related to the project's work:

- Development of affiliated APIs by Internet Archive,
- Migration of all Archive-It institutional users to using the WASAPI data transfer API, including providing technical support and training,
- A "How to Use WASAPI" webinar scheduled for May 2019,
- An open discussion on WASAPI and future work at IIPC's 2019 General Assembly,
- Exploration of additional collaboration between Archive-It and the Archives Unleashed project, extending WASAPI-based collaboration and interoperability, and
- Additional computational research partnerships following on the success of Rutgers University's work in the WASAPI project using the API for research dataset analysis.

Key Project Resources
https://github.com/WASAPI-Community/data-transfer-apis
https://archive.org/details/wasapi